Nucleic Acids Research, 2006, Vol. 34, Database issue D257–D260
doi:10.1093/nar/gkj079

SMART 5: domains in the context of genomes and networks

Ivica Letunic, Richard R. Copley[1], Birgit Pils[2], Stefan Pinkert[2], Jörg Schultz[2] and Peer Bork*

EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany, [1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK and [2]Bioinformatik, Biozentrum, Am Hubland, University of Wuerzburg, 97074 Wuerzburg, Germany

Received September 13, 2005; Revised and Accepted October 11, 2005

ABSTRACT

The Simple Modular Architecture Research Tool (SMART) is an online resource (http://smart.embl.de/) used for protein domain identification and the analysis of protein domain architectures. Many new features were implemented to make SMART more accessible to scientists from different fields. The new 'Genomic' mode in SMART makes it easy to analyze domain architectures in completely sequenced genomes. Domain annotation has been updated with a detailed taxonomic breakdown and a prediction of the catalytic activity for 50 SMART domains is now available, based on the presence of essential amino acids. Furthermore, intrinsically disordered protein regions can be identified and displayed. The network context is now displayed in the results page for more than 350 000 proteins, enabling easy analyses of domain interactions.

INTRODUCTION

When the Simple Modular Architecture Research Tool (SMART) database was first made public 8 years ago (1), the current extent of completely sequenced genomes was little more than a dream. In the last few years, the astonishing successes of whole organism approaches to biology are not only limited to sequencing efforts but also include techniques, such as the high-throughput identification of protein–protein interactions, which have created new opportunities and higher expectations for computational approaches to interpreting biological sequences. In the last 2 years, we have been developing new ways of meeting these challenges.

The basic data of SMART are high-quality manually derived alignments of protein domain families. As hidden Markov models (2) these allow us to identify protein domains in sequence databases; these results are stored in a database accessible via a simple web interface (http://smart.embl.de). The data provide a framework for understanding the evolution and function of genes and proteins throughout the living world. Whereas the SMART philosophy has been to include essentially all available protein sequences, we recognize that many users are interested primarily in the biology of a particular organism. Accordingly, we have developed new views more tightly integrated with genome data. These new genome views allow further cross-referencing with protein–protein interaction maps, making SMART an invaluable tool for systems biologists to interpret pathways and networks.

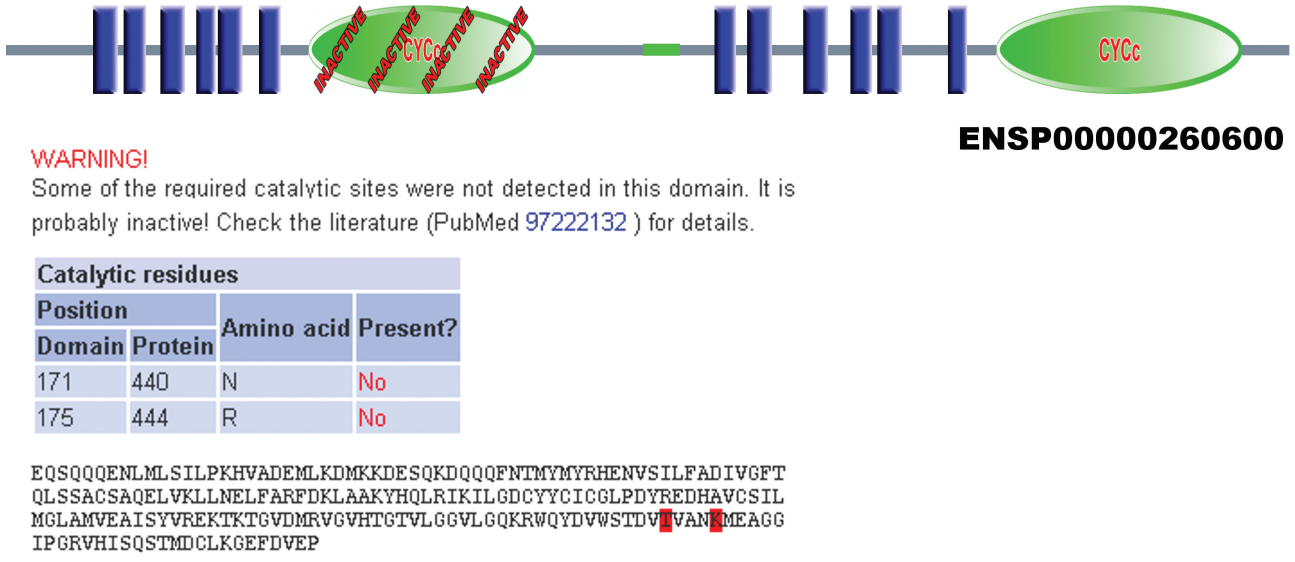
REDUCED PROTEIN DATABASE REDUNDANCY AND 'GENOMIC' MODE

Owing to the nature of our source databases (Swiss-Prot, SP-TrEMBL and Ensembl) (3,4) the protein database in SMART has significant redundancy, even though identical proteins are removed. Different proteins and fragments in the source databases often correspond to the same gene. Users exploring the various domain architectures or interested in domain counts in various genomes are particularly vulnerable to this problem, as the numbers they get are often inflated and unrealistic. To overcome this problem, we extended SMART with a new operating mode, namely 'Genomic' mode. The main difference between normal and genomic mode in SMART is the underlying protein database. In genomic mode, only the proteins from 170 completely sequenced genomes are included (a full list is available at http://smart.embl.de/smart/list_genomes.pl). Swiss-Prot (3) is our main source database of genomic data, together with Ensembl (4) for metazoan genomes. This database has minimal redundancy, and is therefore particularly useful for whole genome studies of domain architectures or single domain distributions.

PREDICTION OF CATALYTIC ACTIVITY

To improve the function prediction for single domains, we annotated essential catalytic sites for all enzymatic domains in SMART. These were extracted from structural reports in the

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de



WARNING!
Some of the required catalytic sites were not detected in this domain. It is probably inactive! Check the literature (PubMed 97222132) for details.

Catalytic residues			
Position		Amino acid	Present?
Domain	Protein		
171	440	N	No
175	444	R	No

```
EQSQQQENLMLSILPKHVADEMLKDMKKDESQKDQQQFNTMYMYRHENVSILFADIVGFT
QLSSACSAQELVKLLNELFARFDKLAAKYHQLRIKILGDCYYCICGLPDYREDHAVCSIL
MGLAMVEAISYVREKTKTGVDMRVGVHTGTVLGGVLGQKRWQYDVWSTDVTVANKMEAGG
IPGRVHISQSTMDCLKGEFDVEP
```

Figure 1. Prediction of catalytic activity in SMART. First guanylyl cyclase domain in human adenylate cyclase type III (ENSP00000260600) is marked as 'inactive' because the two amino acids required for its activity are not present. Domain annotation page shows which amino acids are not detected and gives pointers to the relevant literature.

primary literature, wherever the catalytic mechanism was known (5). Now, protein sequences can be scanned for the presence of important catalytic amino acids (Figure 1). Absence of one of these amino acids very likely results in loss of catalytic activity. Recently, it turned out that many domains homologous to signaling enzymes seem to have lost their catalytic ability, although they are evolutionarily conserved. Instead of a catalytic function these domains appear to play a role in regulatory processes. This trend is especially obvious in the protein tyrosine phosphatase family (5). The inclusion of catalytic amino acid residues in the database will allow a more rapid identification of inactive enzyme homologs in the future.

DOMAIN ARCHITECTURE INVENTION DATING

As a further step from the single domain to the understanding of multi domain proteins, SMART now predicts the taxonomic class, where the concept of a protein, that is its domain architecture, was invented. The domain architecture is defined as the linear order of all SMART domains in the protein sequence. To derive the point of its invention, all proteins with the same domain architecture are mapped onto NCBIs taxonomy (6). The last common ancestor of all organisms containing at least one protein with the domain architecture is defined as the point of its origin. From the knowledge on the origin of domain architectures one might infer the distribution and presence of these architectures in not yet or incompletely sequenced genomes. In addition, conclusions on the general function of domain architectures can be drawn.

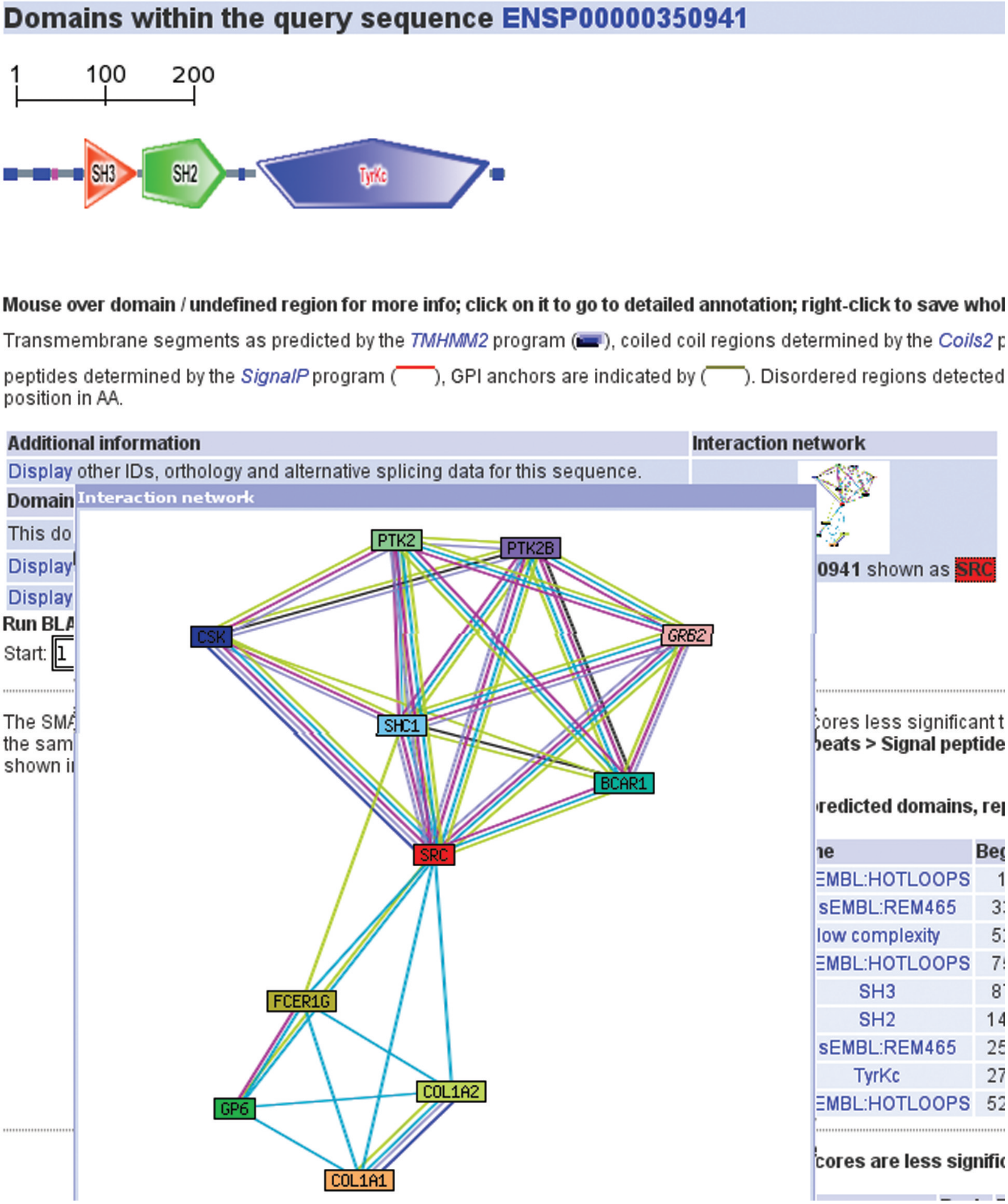
PROTEIN INTERACTION DATA

The latest version of SMART provides information about putative interaction partners for more than 350 000 proteins (Figure 2). This information is imported from the STRING database (7), in which known and predicted protein–protein associations are integrated from a variety of sources. The interactors are shown in SMART in the form of a summary graphic (network); the various types of interaction evidence are depicted as lines of different colors in the network. Clicking on the graphic will launch the STRING website, where the underlying evidence can be studied in detail. The interactions in STRING include physical binding interactions, as well as functional associations, such as membership in a common pathway or process. The data are derived from a variety of sources, including knowledge bases, such as BIND (8), KEGG (9), HPRD (10) and Reactome (11), as well as *in silico* prediction approaches and automated text-mining. STRING aims to improve usability of the interactome by scoring and ranking interaction data (making a confidence estimate on each prediction), as well as by transferring interaction knowledge between model organisms where applicable. SMART and STRING are both cross-referenced through a common set of proteins and genomes, and STRING in turn uses domain information from the SMART server in its pages as well.

NEW DATABASE FEATURES

The core of SMART is a relational database management system (RDBMS) which stores information on SMART domains (1,12). Owing to the exponentially increasing amount of data, many parts of the database access code have been updated or completely rewritten, resulting in greatly improved response times, most noticeably in the domain architecture analysis operations.

SMART database includes the information on domain presence in all proteins in a non-redundant database, now with the added data on the catalytic activity for 50 catalytic domains. All domain architecture analysis results include this information, and domains with missing essential amino acids are overlaid with the word 'inactive' (Figure 1). The domain



Domains within the query sequence ENSP00000350941

Mouse over domain / undefined region for more info; click on it to go to detailed annotation; right-click to save whol

Transmembrane segments as predicted by the *TMHMM2* program (), coiled coil regions determined by the *Coils2* p

peptides determined by the *SignalP* program (), GPI anchors are indicated by (). Disordered regions detected
position in AA.

Additional information	Interaction network
Display other IDs, orthology and alternative splicing data for this sequence.	

Domain

This do

Display

Display

Run BLA

Start: 1

The SMA
the sam
shown i

Interaction network

0941 shown as SRC

cores less significant t
beats > Signal peptide

redicted domains, rep

ne	Beg
EMBL:HOTLOOPS	1
sEMBL:REM465	3
low complexity	5
EMBL:HOTLOOPS	7
SH3	8
SH2	14
sEMBL:REM465	25
TyrKc	27
EMBL:HOTLOOPS	52

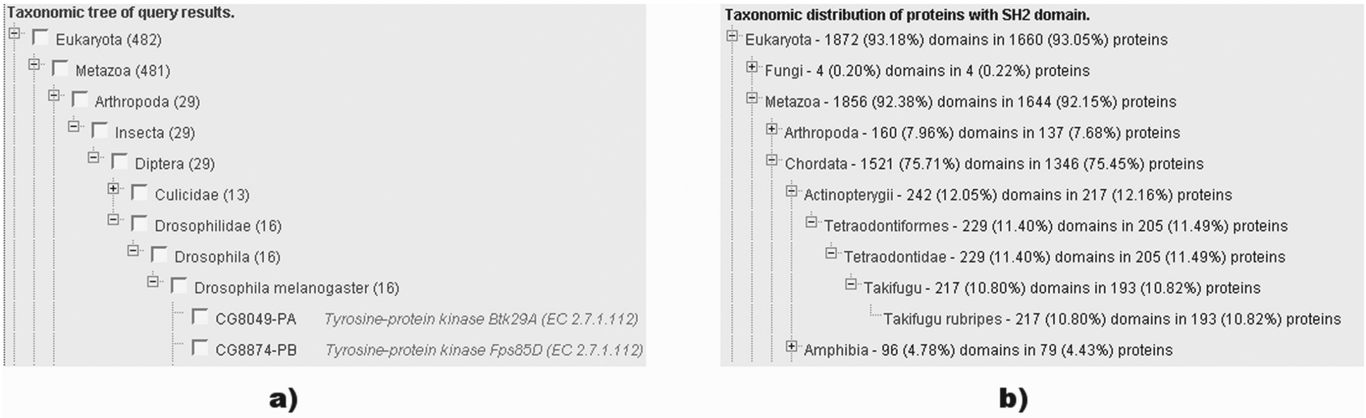
cores are less signific

Figure 2. Interaction networks in SMART. Around 350 000 protein annotation pages include an interaction network in a pop-up window. Networks are linked to the STRING database (http://string.embl.de) which provides the data.

annotation page provides detailed information on which of the required amino acids are missing, and gives pointers to the relevant literature.

NEW ANALYSIS METHODS

DisEMBL [http://dis.embl.de, (13)] predictions of intrinsic protein disorder were included into SMART's analysis methods. DisEMBL is a computational tool for the prediction of disordered/unstructured regions within a protein sequence. Predictions included in SMART are based on missing coordinates in X-ray structure as defined by REMARK465 entries in PDB and the 'Hot loops' method. Hot loops constitute a refined subset of the standard loops/coils as defined by DSSP (14), namely, those loops with a high degree of mobility as determined from C-α temperature factors (B-factors).



Taxonomic tree of query results.

- Eukaryota (482)
  - Metazoa (481)
    - Arthropoda (29)
      - Insecta (29)
        - Diptera (29)
          - Culicidae (13)
          - Drosophilidae (16)
            - Drosophila (16)
              - Drosophila melanogaster (16)
                - CG8049-PA *Tyrosine-protein kinase Btk29A (EC 2.7.1.112)*
                - CG8874-PB *Tyrosine-protein kinase Fps85D (EC 2.7.1.112)*

Taxonomic distribution of proteins with SH2 domain.

- Eukaryota - 1872 (93.18%) domains in 1660 (93.05%) proteins
  - Fungi - 4 (0.20%) domains in 4 (0.22%) proteins
  - Metazoa - 1856 (92.38%) domains in 1644 (92.15%) proteins
    - Arthropoda - 160 (7.96%) domains in 137 (7.68%) proteins
    - Chordata - 1521 (75.71%) domains in 1346 (75.45%) proteins
      - Actinopterygii - 242 (12.05%) domains in 217 (12.16%) proteins
        - Tetraodontiformes - 229 (11.40%) domains in 205 (11.49%) proteins
          - Tetraodontidae - 229 (11.40%) domains in 205 (11.49%) proteins
            - Takifugu - 217 (10.80%) domains in 193 (10.82%) proteins
              - Takifugu rubripes - 217 (10.80%) domains in 193 (10.82%) proteins
      - Amphibia - 96 (4.78%) domains in 79 (4.43%) proteins

Figure 3. Taxonomic trees in SMART. (**a**) Domain architecture query results grouped into a tree. Users can select individual proteins or taxonomic nodes to display. (**b**) Domain annotation pages show detailed domain and protein counts in various taxonomic nodes.

USER INTERFACE IMPROVEMENTS AND TECHNICAL CHANGES

SMART's user interface was completely rewritten and is now fully compliant with the latest web standards, such as XHTML1.0 and CSS2. Users with standards-compliant web browsers can fully enjoy the extra speed and features. Owing to increasing server load, the queuing system was completely rewritten and the hardware greatly expanded resulting in a more stable operation and faster response times.

An important new feature is the introduction of taxonomic trees into SMART. Two primary uses for taxonomic trees in SMART are the grouping of domain architecture query results and the detailed taxonomic distribution of domains now shown on domain annotation pages (Figure 3). The grouping of architecture query results allows users to easily display only proteins from certain species or taxonomic nodes. Taxonomic distribution of proteins on domain annotation pages gives a detailed overview of domain presence in different species and taxa.

ACKNOWLEDGEMENTS

We would like to thank Christian von Mering for providing the interaction network data and STRING links. We are grateful to Rune Linding for helping with the integration of DisEMBL predictions into SMART. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
2. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol*., **235**, 1501–1531.
3. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
4. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
5. Pils,B. and Schultz,J. (2004) Inactive enzyme-homologues find new function in regulatory processes. *J. Mol. Biol*., **340**, 399–404.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
7. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **33**, D433–D437.
8. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.
9. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
10. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*., **32**, D497–D501.
11. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.
12. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*, **32**, D142–D144.
13. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453–1459.
14. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.